DOCUMENT RESUME

ED 185 109 TM 800 122

TITLE The Virginia State Assessment Program.
INSTITUTION Virginia State Dept. of Education, Richmond.
PUB DATE [78]
NOTE 18p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Basic Skills; *Criterion Referenced Tests; Cutting Scores; Elementary Secondary Education; *Equated Scores; Graduation Requirements; Mastery Tests; State Programs; Test Construction; *Testing Programs; *Test Interpretation
IDENTIFIERS Domain Referenced Tests; Rasch Model; Virginia

ABSTRACT

Rasch model statistics and domain-referenced test development methods were used in the Virginia Basic Learning Skills (BLS) testing programs and the Graduation Competencies testing program in order to increase the fairness and effectiveness of those programs. Mastery of each of the 83 objectives in reading and mathematics for grades one through three was certified on the basis of two different criteria. The difficulty level of each of the objectives was determined and located on a scale from 160 to 240. By one criterion, students whose Rasch scaled score was higher than the difficulty level of the objective were deemed competent for that objective. In addition, students were expected to answer three out of the four items correctly for each objective that they had mastered. Instances of discrepancy between these two criteria were to be resolved on the basis of other information, such as the administration of an alternate form of the test. For certification of eligibility for high school graduation, total test scores, expressed in terms of total percent correct, were used. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Va. SDE
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED185109

THE VIRGINIA STATE ASSESSMENT PROGRAM

Recent legislative mandates, known as the Standards of Quality for Public Schools in Virginia (SOQ), have resulted in a more comprehensive testing program than we have previously encountered. Two major components constitute the bulk of our testing efforts under the SOQ; they are the Basic Learning Skills (BLS) testing program and the Graduation Competencies testing program.

During the developmental phases of these programs we conducted an in-depth study of the state of the art in measurement, with particular emphasis upon criterion-referenced testing. As a result of this study, we became quite concerned about the limitations inherent in objectives-referenced tests. Further, the traditional measurement procedures being used to estimate reliability and to equate test forms seemed to be fraught with severe problems. We concluded that it was not in our best interests to develop our testing program exclusively upon traditional measurement theory.

Our search for improved methods to utilize in the development of our tests led us into two areas, domain-referenced testing and the use of the Rasch Model for person measurement, item calibration, and test equating. It was, and still is, our view that these two approaches, coupled with but not replacing traditional measurement theory and test construction methods, have provided us with a testing program which adequately meets our needs.

The focus of this paper will be on our use of the Rasch Model in our test development activities.

TM800 122

Basic Learning Skills

The SOQ clearly call for a testing program in the basic skills which will "...provide the classroom teacher with information to help in assessing the educational needs of individual students." The BLS testing program, when fully implemented, will consist of tests for grades one through six covering reading, mathematics, and communication skills.

Objectives-referenced tests were rejected because of their inability to provide sound results upon which to build an efficacious instructional program. As a result, in order to provide useful information about student needs we a) implemented an item development procedure based upon test specifications (Popham, 1978), and b) developed and implemented a scoring/interpretation system based upon the Rasch Model.

The first operational phase of the BLS testing program (currently in place) consists of six tests measuring 83 objectives in reading and mathematics for grades one through three (attachment). It is expected that students will master these objectives as they pass through the elementary grades. Mastery of each objective is certified by performance on the tests. Mastery of each objective has been judgmentally set at the 75% level. Quite frankly, expediency led to this choice. Information as to the nature of our rationale is beyond the scope of this paper but would be provided upon request.

There are two criteria for the certification of mastery of the objectives, one traditional and one using probability estimates based upon the Rasch Model. It is the interaction between these two criteria that we believe provides the teacher with the maximum diagnostic information possible.

Our interpretation system is based upon two factors, the difficulty (d_i) of an objective and the ability of the student.

The difficulty of each objective was determined after a field test of all items was conducted last spring. We field tested twelve items for each objective (for a total of 996 items) and estimated each item's difficulty using the Rasch Model. The mean difficulty of the twelve items field tested became the difficulty of an objective. After using a simple linear transformation ($10d_i + 200$), each objective was placed on a scale ranging from 160 to 240.

Similarly, ability estimates (the term used in the Virginia program for the Rasch ability estimate is Mastery Level) were calculated for any given total raw score (by form) and transformed on to the same scale as the difficulty estimates. The tables used for the conversion of raw scores to a Mastery Level are devised so that wherever the Mastery Level is equal to the difficulty of an objective the probability is .75 that the student will correctly answer the items measuring the objective. Note that this probability of a correct response is identical to the percent correct for determining mastery of each objective. This probability is derived directly from the Rasch Model algorithm:

$$P_{vi} = \frac{e^{(b-d)}}{1+e^{(b-d)}}$$

where P_{vi} = the probability of a correct response
b = the person's ability,
d = the item's (objective's) difficulty, and
e = 2.72 (base e)

The probability of correctly answering the items measuring an objective increases as the positive difference between a student's Mastery Level and an objective's difficulty increases. As this difference gets smaller, or becomes increasingly negative, the probability of correctly answering the items decreases. For example, if a student's Mastery Level is 200

and an objective's difficulty is 190 (difference is 200-190 = 10), the probability of correctly answering the items measuring the objective (P_{vi}) is .89. If the Mastery Level remains the same and the difficulty of an objective is 211 (200-211 = -11), the probability decreases to .50. The operational test forms, in most cases, contain four items per objective. The mean difficulty of these four items is equal to the mean difficulty of the twelve item set (pool) from which they are drawn. Further, none of the four items deviate more than one standard error from the mean difficulty of the twelve item pool.

A brief excerpt from our User's Guide follows, describing the scoring and interpretation of the tests.

5.3 - Interpretation

It was stated previously that a scoring system had been adopted that was an extension of traditional criterion-referenced scoring. This system has been developed to provide maximum information to the teacher about individual student's strengths and weaknesses.

To illustrate the increased diagnostic power of the BLS scoring system, it is useful to examine the results obtained from traditional CRTs and the types of interpretations made possible by them. In traditional CRTs, students receive a raw score indicating the number of correct responses for any given objective. The interpretation of this score is that if the student answers enough of the questions correctly; e.g., 75% or more, than he/she is a "master" of that objective. If the student answers correctly fewer than 75% of the items associated with an objective, then the student is classified as a "non-master". Instructional effort is then placed upon re-teaching those objectives to students which they have not mastered, with little or no subsequent attention given to mastered objectives.

All measurement of student achievement is, to some degrees, imprecise. Thus, it is possible that a student may be certified by a test to be a master of an objective when he/she really is not a master (false-positive). Similarly, a student may be certified as a non-master when in fact the student actually does possess the skill being measured (false-negative). The more imprecise a test, the more likely that students will be incorrectly categorized as mastering or not mastering the objective of interest. What are the instructional implications of these incorrect decisions? In the case of a false-positive situation, the teacher may not remediate a student who actually needs help in the skill. In the case of a false-negative situation, the teacher may commit resources (time, materials, etc.) to the

remediation of a student who does not really need it. The scoring system being used for the BLS should more precisely classify individual students as masters or non-masters, thereby increasing efficiency in allocating resources in remedial efforts.

The two principal components for determining mastery are, the number items answered correctly for each objective and, the line on the Student Profile representing the MASTERY LEVEL. When is it appropriate to use one or both of these criteria?

For the BLS tests, a student can be certified as mastering an objective when:

1. The objective falls AT or BELOW the line representing the student's MASTERY LEVEL on the Student Profile, and/or
2. The student has answered AT LEAST three items correctly for any objective.

To see how this may work in actual practice look again at SP 3. Applying criterion 1 above to SP 3 we see that the student has not mastered objectives 8.1, 4.2, 5.2, 9.1, and 1.13 because they all fall above the line representing the student's MASTERY LEVEL. The student can be certified as mastering the remaining objectives because they all fall AT or BELOW the line.

Next using criterion 2 above, the student would also be certified as a non-master of the same five objectives identified as falling above the mastery level line. Using this example, both criteria have provided the same mastery decisions.

These results suggest that the student should be given instruction in the five non-mastered objectives, with relatively little further attention given to the objectives which the student has mastered.

Prior to completing this section on interpreting test results, it would be well to explain the relationship between the placement of the objectives on the Student Profile and the MASTERY LEVEL. Each objective is placed on a scale which ranges from 160 to 240. The scale represents the difficulty of each objective, low numbers indicating easy objectives and high numbers indicating difficult ones. The difficulty values of the objectives were calculated after field-testing and then were placed on the appropriate point of the scale. The easier objectives are placed at the bottom of the scale and the more difficult are placed at the top of the scale. Thus, the placement of the objectives on the profile represents how easy or difficult they are compared with each other.

Similarly, each total raw score on the test has been converted to a MASTERY LEVEL score. This places both the difficulty of the objectives and the student's MASTERY LEVEL on the same scale. As a result, where the difficulty of an objective and a student's MASTERY LEVEL are identical, it is fairly certain (75% certain) that the student has mastered the objective. A convenient way to think of the relationship between the difficulty of an objective and a student's MASTERY LEVEL is to think of the latter as an "expected" level of performance. Using SP 3 as an example it might be said, "Given a MASTERY LEVEL of 200 it is expected that a student should master all of the objectives which have a difficulty equal to or less than 200." Using the criterion of at least three correct responses per objective as one of two indicators of mastery, this expectation is confirmed.

There may, however, be instances when there is a discrepancy between the two criteria for determining mastery of an objective. Sample Profile Number 4 (SP 4) is an example of such a discrepancy.

VIRGINIA BASIC LEARNING SKILLS
STUDENT PROFILE — READING

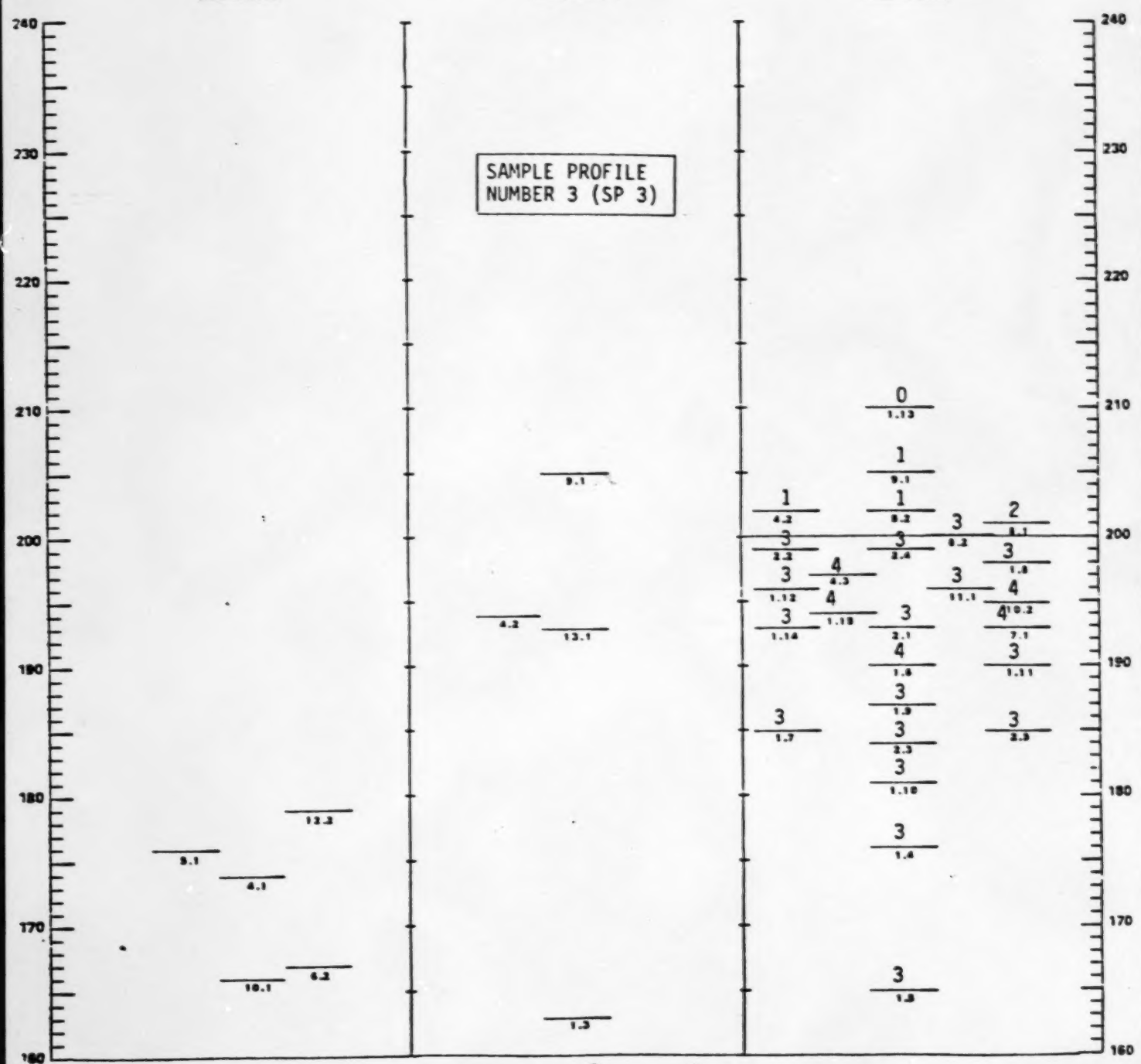
R0878

NAME SAMPLE NUMBER 3

RAW SCORE 73 MASTERY LEVEL 200 DATE 10/13/78 FORM/LEVEL A-1 / 3

LEVEL 1 LEVEL 2 LEVEL 3

SAMPLE PROFILE NUMBER 3 (SP 3)

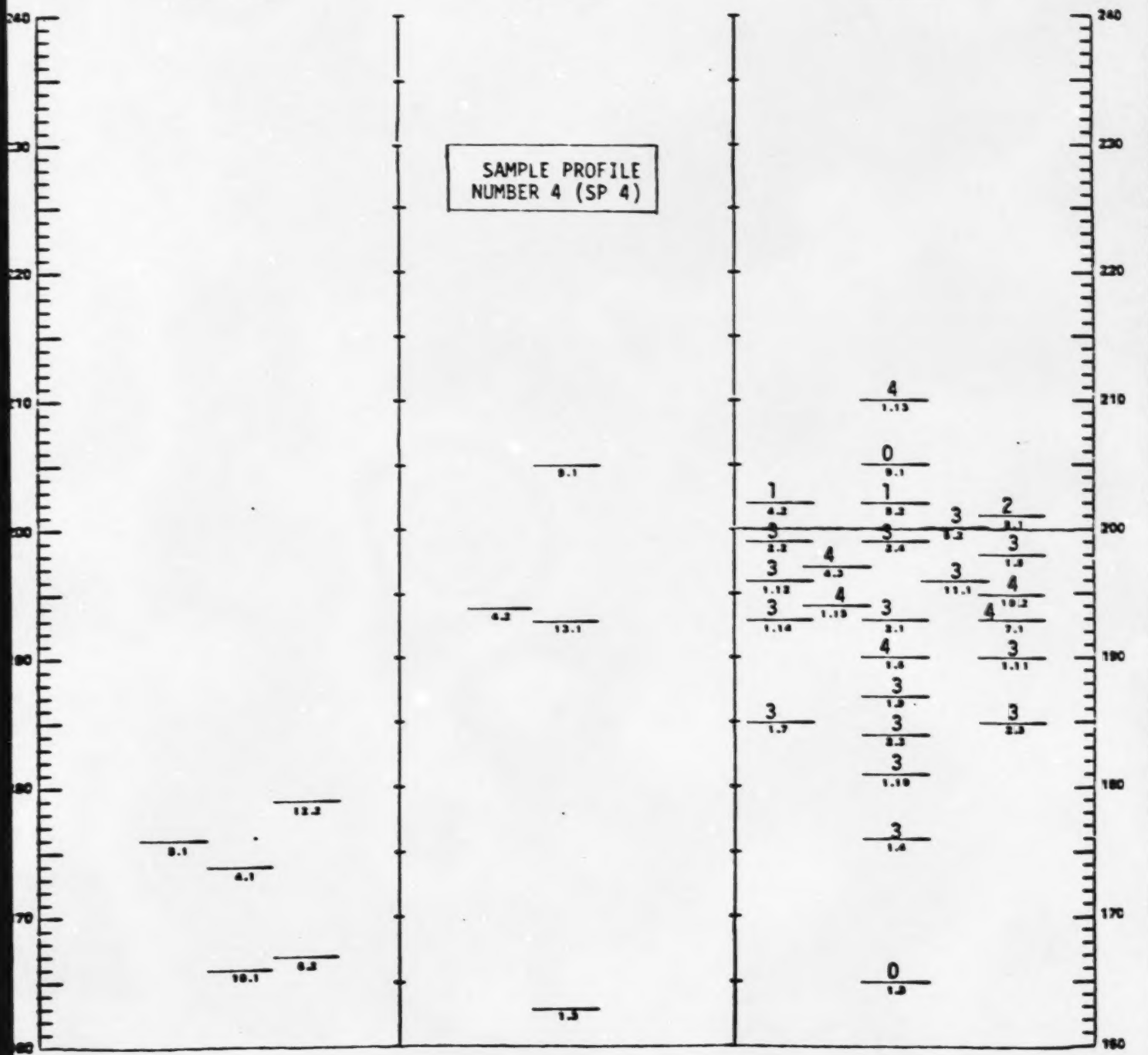


STUDENT PROFILE — READING

NAME SAMPLE NUMBER 4

RAW SCORE 73 MASTERY LEVEL 200 DATE 10/13/78 FORM/LEVEL A-1 / 3

LEVEL 1 LEVEL 2 LEVEL 3



In this example (as in SP 1-3) the student has correctly answered 73 items and the MASTERY LEVEL is therefore 200. There is, however, some apparently conflicting information about whether the student has mastered objectives 1.13 and 1.5. Using the MASTERY LEVEL as the criterion of mastery the teacher would determine that the student should not have mastered objective 1.13 and should have mastered 1.5. Yet, the other criterion of mastery, the number of correct responses per objective, suggests that the student <u>has</u> mastered objective 1.13 and has <u>not</u> mastered 1.5. This is a surprise and, in fact, the further below the MASTERY LEVEL non-mastery of an objective occurs and the further above the MASTERY LEVEL mastery occurs, the more surprising is the discrepancy.

Given that these "surprises" have taken place, how do you proceed? Keep in mind that if <u>only</u> the criterion of at least three correct answers per objective we would not remediate the student on objective 1.13. Also, we would most likely begin a remedial sequence for this student on objective 1.5 (and other students who had apparently not mastered the objective). Would this be the "correct" action to take?

Since we would not expect a student to master an objective which has a higher difficulty level than his/her mastery level, we may wish further investigation to confirm or refute what the test results seem to indicate. Further confirmation is needed because other factors may be responsible for this unexpected mastery. Some of these factors may be:

1. the student may have copied from another student's paper;
2. the student may have been a very lucky guesser;
3. the student may have had some special "coaching" in the skill (objective 1.13), in which case the skill may not really be learned; or
4. the presumed hierarchy (based upon each objective's difficulty) is a function of instruction and not development. Therefore, the order of learning can be shifted and has been for this student.

There could very well be other factors which produced the discrepancy, but the important point is that we must be a bit surer about the student's mastery status before we finally certify him/her as a master. In other words, we need to know if the test has incorrectly classified the student as a master when he/she really isn't (false-positive). To resolve this discrepancy there are some procedures you may wish to follow in order to make a final determination of mastery:

1. Administer the Form B test for objective 1.13 to the student: If the student achieves mastery on this test then he/she can be classified as mastering the objective. If the student does not achieve mastery on the Form B test you may wish to consider instructional help for the student, after which the Form B could be re-administered for final mastery determination.

2. Evaluate the classwork which the student has completed which is similar to the objective in question. If the student's performance on the classwork is questionable, supply remedial support and re-evaluate the student with Form B when the remedial help has been concluded. If the classwork is acceptable, re-evaluate the student using Form B and certify mastery.

3. Any appropriate combination of the above.

The second anamolous characteristic presented by SP 4 is the apparent non-mastery of a very easy objective, objective 1.5. Given a MASTERY LEVEL of 200, mastery of objective 1.5 would be expected, and it is very surprising that the student has not mastered it. We need to determine whether or not the test has incorrectly classified the student as a non-master when in fact he/she may be a master. Some of the reasons this discrepancy may have taken place are:

1. there may have been a disturbance within the classroom during testing which diverted the student's attention;

2. the student may not have been present in the classroom when the skill had been taught;

3. the teacher may have been absent and the substitute failed to teach the skill or was ineffective in his/her instruction;

4. the student may not have understood the test directions, or

5. the items may have been too easy for the student, so due to boredom or fatigue the student may have been careless in marking answers.

A further determination of the student's true status as either a master or non-master is indicated prior to the initiation of a complete remedial sequence. It is probable that when a student fails to master an objective which is far below the MASTERY LEVEL that the student could acquire the skill with a minimum of instruction and that an exhaustive instructional sequence would be unnecessary. The procedures described earlier for determining the student's status when he/she masters an objective which is too difficult (well above the student's MASTERY LEVEL) would be appropriate here, also.

End of Excerpt

One point which has not received enough attention is the selection of test items for alternate forms of criterion-referenced tests. It has been our experience that, in most cases, when alternate test forms are constructed the major concern is to ensure that the items measuring an objective have the same mean difficulty and discrimination on each form. The obvious goal here is to consistently classify students as masters or non-masters on each form.

While "equating" tests using the mean item difficulty has merit, it also has severe limitations. These limitations include using p-values for estimating the difficulty of the items (which are sample dependent) and not paying sufficient attention to the dispersion of the difficulty of the items. For example, the following two sets of test items could be constructed from our item bank. Each set measures the same objective and could be used on alternate test forms.

Item set 1 (Rasch item difficulties): 197, 198, 199, 201, 202, 203

Item set 2 (Rasch item difficulties): 195, 197, 199, 201, 203, 205

Note: A difficulty of 200 may represent a p-value of .50

The mean difficulty of each item set is 200, yet if these item sets were used on alternate test forms they would lead to a different number of children classified as mastering the objective. By way of illustration, suppose the criterion for mastery was correctly answering five of the six items. A correct answer to five of the six items in item set 1 represents a Rasch ability (Mastery Level) of 216. A correct answer to five of six items in set 2 represents a Rasch ability of 220. Thus, based on the distribution of ability in Virginia, on our third grade math test this would represent a group of approximately 2,000 children who would be classified as mastering the objective if they encountered item set 1 but as not mastering the objective if they encountered item set 2.

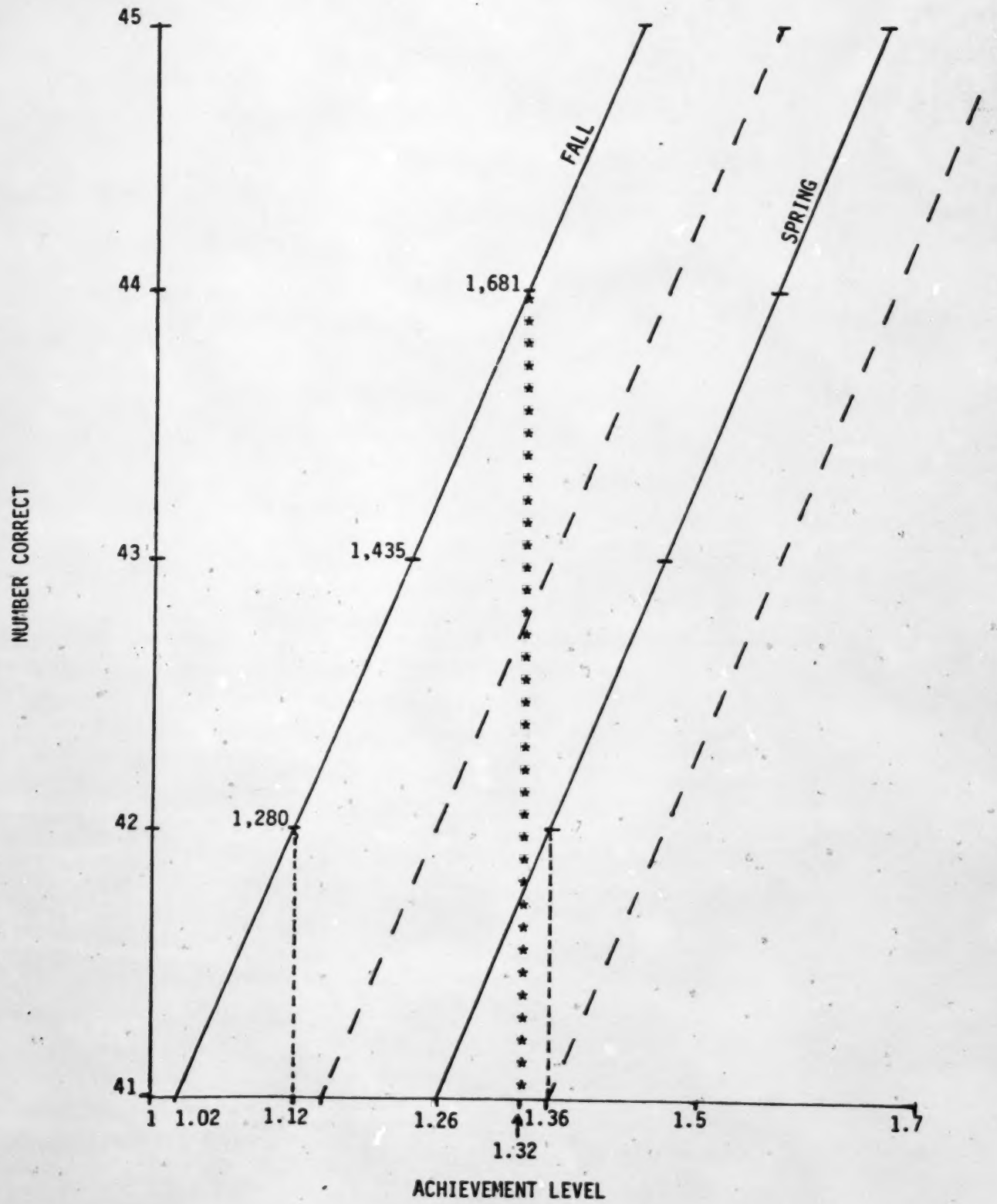
Generally, attempts to avoid variations in classification due to differences in difficulty have focused upon matching items measuring the same objective on alternate test forms on a p-value by p-value basis. There are two major deficiencies to this approach. First, the sample dependency of p-values may lead to problems in obtaining stable difficulty estimates. Second, an extremely large item bank may be needed in order to provide a sufficient number of test items to guarantee identical mean difficulty and dispersion on any alternate forms. Utilization of the Rasch Model in the equating of test forms ameliorates these problems by providing stable difficulty estimates and by assuring that the score needed for mastery is the same between test forms.

Graduation Competency Tests

The preceding discussion has emphasized the importance of equating forms for criterion-referenced test (CRT's) on an objective by objective basis. The deficiencies of traditional approaches to item calibration, person measurement, and test equating which were discussed for CRT's are also applicable to the tests being used/developed for graduation competency programs. The major exception is that total test scores, expressed in terms of total percent correct, rather than scores on an objective by objective basis, are being used for the certification of graduation eligibility.

Again using data from the Virginia program, it has been determined that the spring form of our competency test in reading is slightly more difficult than the fall form (Table I).

Table I
Disparity in Difficulty Between Two Forms of
the Reading Graduation Test (Virginia)



As Table I illustrates, the ability estimate for 42 correct responses (our 70% cut-score) on the fall form is less than the ability estimate for the same number correct on the spring form. In fact, students who come to our spring form, with sufficient knowledge/skills to obtain scores of 42-44 on the fall form, may fail the spring form simply due to the difference in difficuly between the forms. A total of 4,396 students scored 42, 43, or 44 on the fall form. If this distribution remains equivalent for the spring form there is a distinct possibility that over 4,000 students may fail the spring form as a function of its difficulty and not a lack of "competency" on their part. This is an intolerable situation.

Difference in difficulty between test forms is not a problem unique to Virginia. Some of the alternatives which may be available to ameliorate differences in difficulty between forms are:

1. If possible, pre-equate alternate forms in the area of the cut-score so that ability estimates in this range are identical between forms. Ideally, the tests would be equated all along the raw-score range. If tests are equated only in the range around the cut-score individual competencies measured by the test may in fact be easier or more difficult even thoug the tests are equated at the cut-score. One result of this may be that if on subsequent testings one competency is more difficult than on prior testings, and fewer students are certified as passing that competency, then erroneous interpretation may be made that students aren't "doing as well" as they did before. If possible, tests should be equated all along the score range.

2. The notion of using pre-equating methods based upon matching the mean, variance, and discrimination among test items on alternate test forms is not by any means a new idea. Using a scaled or standard score as the cut-score may be, particularly on graduation competency tests. Essentially what this would mean is to use a procedure where an arbitrary mean and standard deviation is selected and then transforming the Rasch ability estimates to a new scaled cut-score. For example, on the Virginia test the ability estimate could be transformed (using the results of the initial testing) in this fashion; 10 (bv) + 59, where bv is the ability estimate for any raw score. The scaled score representing 70% correct would then be 10 (1.12) + 59= 70. Thus, in order to pass any subsequent test the students would need to obtain a scaled score of 70. Using the results of our pre-equating in the construction of our spring form would lead to a passing score of 65% correct, i.e., the ability estimate for 39 items correct

on the more difficult spring form is 1.1, or a scaled score of 70. However, due to the fact that there is some error in our estimate (± 1 SE is 67 to 73 scale points) it is recommended that, at least for the first administration of an alternate form, a post-equating procedure be utilized to evaluate the integrity of the prior estimates.

3. A possible problem with this approach may be that the percent correct may vary from form to form due to differences in the difficulty of the test forms. Explaining these differences to the public may be difficult as the charge may be made that the scores are being artificially tampered with. One course of action may be to include additional items on subsequent forms. These additional items could be included as "experimental" items. After the results of the testing were analyzed, items could be deleted in a manner such that the percent of correct answers remaining would remain the same as previous testing and represent the same Rasch ability as earlier forms.